Frequently asked questions (FAQs) regarding evidence synthesis

This is an evolving document developed in response to similar queries received by the NICE Technical Support Unit (TSU). Questions have been grouped under similar topics. New questions and answers will be added on an ongoing basis.

Regarding WinBUGS

1. How does one know when convergence has been achieved?

   There are no built-in functions available in WinBUGS to let you know exactly when convergence is achieved. Convergence may be assessed by visually inspecting the mixing of chains (you need to run the model with at least 2 chains) in the history plots. These plots are obtained by clicking the "history" button in the Sample Monitor Tool. Plots of the Brooks-Gelman-Rubin diagnostic statistic across each iteration may also be inspected. These plots are obtained by clicking the "bgr diag" button in the Sample Monitor Tool. In a converged model, all lines will be roughly horizontal and stable, with the red line approximately located at a value of 1 on the y-axis.

   A suggested approach is to inspect the history and bgr diagnostic plots every 10,000 iterations for the d[k], sd (when running a random effects model), and totresdev nodes until the convergence is achieved. Then, to record the results, sample double the number of iterations it took to reach satisfactory convergence from the posterior distributions (e.g., if it took 30,000 iterations for satisfactory convergence, base the results on a further 60,000 samples).

2. What does it mean when a history plot consists of a single horizontal line (there is no variability in the samples)?

   This means the parameter has not been estimated. This problem occurs rarely and can usually be mitigated by running the model in OpenBUGS. We have not found an explanation for this issue - it appears to be a rare numerical error in WinBUGS. This is another reason why it is important to inspect the history plots of the treatment effects.

Regarding random effects models

3. When is a random effects model preferred over a fixed effect model?

   A random effects model is preferred over a fixed effect model when there is evidence of heterogeneity between treatment effects estimated by trials making the same comparison. To justify the choice of a random effects model in a Bayesian analysis, the recommended approach is to compare the fixed and random effects models' residual deviance and DIC statistics [1]. Lower values are preferred and typically differences of 3-5 points are considered meaningful [2]. Furthermore, the residual deviance can be used to assess the goodness of fit of each model. In a well-fitting model, the posterior mean residual deviance should be close to the number of data points in the network (each study arm contributes 1 data point) [2].

   Even when there is no statistical evidence of heterogeneity, a random effects model might be preferred if the clinical opinion is that a degree of between-trial variation (in the true treatment effect) would be expected *a priori*.

4. What are the concerns surrounding the estimation of between-study heterogeneity in random effects models?

   There may be instances where a random effects model may appear to provide a better fit over a fixed effect model, but there is not enough evidence to estimate the between-study heterogeneity. Random effects models are routinely fitted with a vague prior distribution set on the between-study heterogeneity parameter and so the estimation of this parameter is largely dependent on the data. In the absence of large numbers of trials for at least one comparison (at least four or five trials, has been suggested as a minimum [3], although three may be sufficient), the posterior distribution of the between-study standard deviation (σ) will be poorly identified and likely to include values that, on reflection, are implausibly high or, possibly, implausibly low. See Item 5 for advice on assessing whether there is enough evidence to estimate the between-study heterogeneity.

5. How can one assess whether the between-study heterogeneity has been adequately estimated?

   If a vague prior, such as Uniform(0,5), is used, the posterior distribution of σ should always be inspected to ensure that 1) it is sufficiently different from the prior as this would otherwise indicate that the prior is dominating the data and no posterior updating has taken place (Figure 1) and 2) there are no "spikes" in the posterior distribution at 0 (Figure 2). It is not uncommon, particularly when data is sparse, that MCMC sampling can "get stuck" at σ = 0, leading to spikes in the posterior distribution of both σ and the treatment effect parameters relative to the reference treatment (d_{1k}).

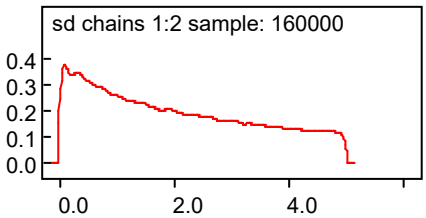


Figure 1: Example of a posterior distribution of between-study σ dominated by its prior distribution, Uniform(0,5). Note it is quite possible for σ to be any value between 0 and 5.

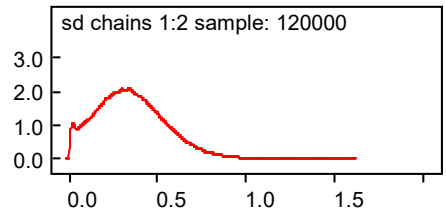


Figure 2: Example of a posterior distribution of between-study σ with a spike near 0.

6. What are my options when the between-study heterogeneity has not been adequately estimated?

   Apart from a fixed effect model, four alternatives may be useful when there is insufficient data to adequately estimate the between-study heterogeneity. The first is the use of external data [4]. If there is insufficient data in the meta-analysis, it may be reasonable to use an estimate for σ from a larger meta-analysis on the same trial outcome involving a similar treatment for a similar condition. The posterior distribution from such an analysis could be used to approximate an informative prior distribution. Alternatively, prior distributions derived from large numbers of meta-analyses can be used, with the appropriate prior distribution chosen depending on outcome type and the type of treatments being compared [5-7]. Thirdly, an informative prior distribution can be elicited from a clinician who knows the field. This can be done by posing the question in this way. "Suppose we accept that different trials, even if infinitely large, can produce different effect sizes. If the average effect was an odds ratio of 1.8 [choose a plausible average], what do you think an extremely high and an extremely low effect would be, in a very large trial?" Based on the answer to this it should be possible, by trial and error, to construct an informative Gamma prior distribution for $1/\sigma^2$, or a Normal prior distribution for σ, subject to σ > 0 (half-Normal) [1]. Finally, a combination of the previous two approaches may be used,

where expert opinion is elicited to truncate a wide informative prior distribution, e.g., [5-7], to a more plausible range of values [8].

7. How should the between-study SD be interpreted?

   The between-study SD should be interpreted in the context of the scale of measurement. If the relative effects are on the SMD or LogOR or LogRR scales, then values of 0.5 or even 0.25 for the between-study SD are large. If you are working on a continuous scale, modelling mean differences, then it depends on the units for the mean difference as to what is considered large. It is helpful to think about 4xSD which is the range for a 95% interval from the distribution of study estimates. Are mean differences on each end of this range (pooled MD±2SD) implausibly far from each other?

Regarding treatment definitions

8. How should different doses of the same drug be incorporated in a network?

   Classifying different doses of the same drug (e.g., 'drug A 40 mg' and 'drug A 80 mg') as separate interventions allows a committee to be more explicit in their recommendations, so they can recommend a drug at a specific dose, rather than just the drug. Having more specific definitions of the intervention may lead to sparsity in the network, risking potential disconnections in the network. We may overcome this by fitting a class effects model, which will still produce estimates for each specific intervention (drug and dose).

   Classifying different doses of the same drug as the same intervention (e.g., 'drug A 40 mg' and 'drug A 80 mg' are categorised as 'drug A') is okay if there is evidence supporting no added benefit of higher doses (e.g., a dose-response curve plateaus at the smallest dose). If this is not the case, then any variability between these treatments' effects will inform the between-study heterogeneity. We typically assume a common between-study variance throughout the network, so it is important to consider whether the variability between the effects of the different doses is typical of the between-study variability in the remainder of the network. Lumping doses together in this way is likely to increase heterogeneity.

9. Suppose a study compares similar interventions, which are only distinguished by a component that is not of interest to the decision (e.g., A + B vs. A, where A is of interest, and B is not). Is it helpful to include this evidence in the network?

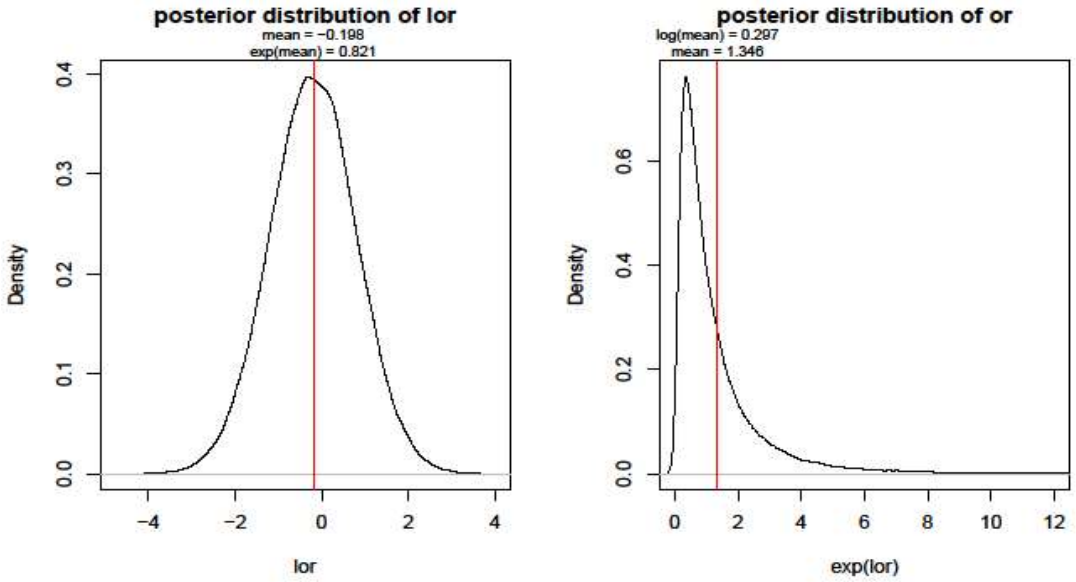
   In these cases, it is important to think about what the relative effect from the study is estimating. Assuming no interaction between A and B, then the relative effect provides information on the effect of B. If we are not interested in B, then this study provides no relevant information, and we recommend not including this study. If a class model was fitted and we categorised interventions 'A' and 'A + B' to fall under class 'A', then this evidence would inform the within-class variability. However, this is variation due to a factor that we are not interested in and it may not be exchangeable with the variations that we are interested in. So, again, we recommend not including this evidence.

Regarding the reporting of results

10. What summary statistics should be reported to describe the posterior distributions of the relative treatment effects?

    We recommend reporting the median and 95% credible intervals (CrIs) at a minimum. The median and 95% CrIs are preferred over the mean and standard deviation when dealing with treatment effects that are synthesized on a transformed scale (e.g., log-odds ratios (LOR)). This

is because one cannot obtain the posterior mean LOR by taking the log of the posterior mean OR as they are not equal. Whilst the LOR distribution is approximately normal, the OR distribution is skewed (with a long tail towards larger values). Since the mean and SD are not robust to extreme values, the log of the posterior mean and SD of the ORs will typically be larger than the mean and SD of the LORs. The figure below demonstrates this.



(The log transformation does not "undo" the influence of the extreme values on mean and SD as this is just a conversion of scale that does not perfectly capture the random variability of the extreme values). If one wanted to convert LOR posterior summaries to OR posterior summaries and vice versa through the exponential or log transformations, then the median and CrIs for either distribution should be reported, as these statistics are invariant to skewness and extreme values.

In practice, the mean and median of the LOR are usually very similar so that exp(mean(LOR)) will approximate exp(median(LOR)) and therefore the median(OR). However, to be strictly correct the exp(median(LOR)) or median(OR) and credible limits should be reported.

11. To transform the estimated pooled log odds ratios from the NMA into relative risks for reporting results, what evidence should be used to inform the baseline probability?

The evidence used to inform the baseline probability of a given treatment should be the same as the evidence used to inform the baseline model used to obtain the absolute treatment effects in the economic model. See item 13 below.

Regarding parameters for economic models

12. How are the absolute treatment effects obtained from an NMA model?

Absolute treatment effects are obtained by using an appropriate mathematical operation (e.g., addition, multiplication) to combine the baseline probability of event and relative treatment effects on the same scale. For example, in the case of a binary outcome, the odds ratios may be added to the odds of an event (informed by the baseline probability) for the baseline treatment on the log scale. This returns the absolute probability of an event for each treatment on the logit

scale. Absolute effects on the probability scale may then be obtained by using the "expit" function, which inverts the logit.

13. Which evidence should be used to inform the baseline model?

    The baseline model is used to estimate the outcome on an absolute scale (eg mean, log-odds, probability) for a reference treatment. Ideally, this is informed by a large observational study conducted in a population representative of the population for which the decision (recommendation) is to be made. In the absence of such data, evidence from RCTs including the reference treatment may be used. RCTs used for the baseline model should be chosen to be representative of a contemporary population for which the recommendations are being made. For this reason there may be only a subset of the RCTs that included the reference treatment that are included in the baseline model. The reference treatment does not necessarily have to be placebo or another version of control, it can be any treatment that is connected in the network. It doesn't even need to be the reference treatment used in the economic model. The main thing is that it is a treatment where there is good evidence in the relevant population with which to estimate the baseline model, and where the included studies have good internal and external validity. Pooling the arms containing the same treatment across all studies in the NMA is only appropriate if these studies' populations are representative of the population for which a decision is to be made.

14. What are the concerns about fitting a random effects baseline model?

    Checking to see if there is enough evidence to estimate the between-study SD in a random effects model (see Item 5) is also important when estimating the baseline model. Two studies are not sufficient to estimate the between-study SD. If there is some evidence of heterogeneity between the effects estimated in a random effects model, this should serve as a flag to re-evaluate the external validity of the studies with respect to the target population for the recommendations.

    If all studies are deemed valid for the baseline model, and there is between-study heterogeneity, then the predictive distribution of the baseline effect, rather than the posterior distribution, should be taken from the baseline model and used together with the relative effects from the NMA to calculate the absolute probabilities for use in the economic model. Be careful though, as large between-study variability might propagate large uncertainty into the economic model, and the results may be meaningless. It is preferable to use stricter inclusion criteria to reduce heterogeneity and make the predictions more relevant to the target population for the recommendations.

15. What are my options when the between-study heterogeneity has not been adequately estimated in a random effects model for the baseline effect?

    If there are a small number of studies (e.g., 2) that have good internal and external validity, but their estimated effects are heterogeneous, then we would recommend selecting one study to inform a base-case analysis, and using the other study/ies instead may be explored in a sensitivity analysis. Assess whether the recommendations would change based on which study has been selected to inform the baseline effect.

    Note that if there is only 1 study, then a random effects model should not be fitted as there is no between-study heterogeneity. In fact, a fixed effect model does not need to be fitted, as there are no studies to pool. The baseline effect can be calculated manually using the appropriate formula, which should produce an estimate similar to the estimate produced by a fixed effect model.

16. Since the posterior mean of the log-odds ratio does not equal the log of the posterior mean of the odds ratio (i.e., mean(LOR) ≠ log(mean(OR))), which parameter should be inputted into the economic model to represent the absolute effect?

    There is no "right answer". You can use either the mean or the median, and they both have advantages/disadvantages:

    1) The mean of the probabilities is appealing because it is the expected probability, and for models linear in probability, this translates into expected net benefit (which is what you want). The problem with it is that for non-linear models then the expected probability doesn't translate into expected net benefit.

    2) The median has the advantage that it remains the median after transformation (as long as the transformation is monotonic). So, the median probability would give the median net benefit, regardless of the model, all else being fixed. But it doesn't have the interpretation of an expected net benefit.

    Nevertheless, this issue can be avoided altogether by using a probabilistic decision model, as the simulated values of the absolute probabilities may be inputted directly into the economic model. Alternatively, you may use the simulations of the baseline log-odds, as well as the simulations of the LORs and ORs, provided that the appropriate conversions are made to combine them into a probability. You should get the same results for both approaches.

Regarding TSU involvement

17. What is required for a QA by the TSU?

    We will need:

    1) The final version of the WinBUGS code used for all NMAs in the guideline.
    2) The final version of any text related to the NMAs including the methods and results.
    3) A description of how the NMA results were adapted as inputs into the economic model (if applicable).
    4) The recommendations made in the guideline.

References

1. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials. 2011; last updated September 2016. Available from http://scharr.dept.shef.ac.uk/nicedsu/technical-support-documents/evidence-synthesis-tsd-series/.

2. Spiegelhalter DJ, Best,NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*. 2002. 64(4):583-616.

3. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis.* 2006. 1:515-533.

4. Higgins JPT, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Statistics In Medicine.* 1996. 15:2733-2749.

5. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International Journal of Epidemiology.* 2012. 41:818-827.

6. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine*. 2015. 34: 984-998.

7. Rhodes KM, Turner RM, Higgins JPT. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology.* 2015. 68:52-60.

8. Ren S, Oakley JE, Stevens JW. Evidence synthesis for health technology assessment with limited studies. Poster presented at: ISPOR 20th Annual European Congress; 2017 Nov 8; Glasgow, Scotland.